

A Permutation Approach for Selecting the Penalty Parameter in Penalized Model Selection

Jeremy Sabourin¹, William Valdar^{1,2}, and Andrew Nobel³ *

¹*Department of Genetics, University of North Carolina at Chapel Hill, North Carolina*

²*Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, North Carolina*

³*Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, North Carolina*

April 9, 2014

Abstract

We describe a simple, efficient, permutation based procedure for selecting the penalty parameter in the LASSO. The procedure, which is intended for applications where variable selection is the primary focus, can be applied in a variety of structural settings, including generalized linear models. We briefly discuss connections between permutation selection and existing theory for the LASSO. In addition, we present a simulation study and an analysis of three real data sets in which permutation selection is compared with cross-validation (CV), the Bayesian information criterion (BIC), and a selection method based on recently developed testing procedures for the LASSO.

1 Introduction

The analysis of high dimensional data, in which the number of measured predictors is large and can exceed the number of samples, is an important and common problem in statistical applications. When samples are accompanied by a real or categorical response, data analysis typically includes model fitting with the aim of doing prediction or variable selection, or both. The goal of prediction is to derive a rule capable of accurately predicting the response of a new, unlabeled sample. The goal of variable selection is to select a (small) subset of the measured predictors whose individual or coordinated activity is significantly related to the response. In both cases, it is common to assume that the observed data arise from an underlying model that is sparse, in the sense that only a small subset of the predictors are related to the response. Whether sparsity is assumed, or viewed as a desirable feature of a model, analysis of high dimensional data is often carried out by penalized methods that produce models in which a relatively small subset of the available predictors are included. Popular penalized methods include the LASSO (Tibshirani, 1996), its numerous variations, and SCAD (Fan and Li, 2001). In what follows, we focus our attention on the LASSO.

The LASSO and its variants require specification of a penalty/tuning parameter that controls the tradeoff between model fit and model size. Specification of this parameter is necessary to fully determine the selected model, and each value results in different coefficient estimates; selecting a suitable value of the penalty

*Correspondence to: Andrew Nobel, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7265, USA. E-mail: nobel@email.unc.edu

parameter is therefore a vital part of model fitting. At present, the most widely used procedures for selecting the penalty parameter in the LASSO are the Bayesian Information Criterion (BIC) and cross validation (CV). Cross validation, which is based on out of sample prediction, is a natural choice when the goal of model fitting is prediction. CV has some empirical and theoretical justification (see, for example, Bühlmann’s comments in Tibshirani, 2011), but in high-dimensional settings CV tends to be too conservative (Feng and Yu, 2013). BIC, although designed primarily for variable selection, is based in part on how well the selected model fits the data, making it a reasonable choice for either prediction or variable selection. In practice, BIC is a popular method, but has no theoretical justification for variable selection with the LASSO (Bühlmann and van de Geer, 2011).

In this paper we address the problem of selecting an appropriate penalty parameter for the LASSO when the primary goal of model fitting is variable selection. We investigate a simple, permutation based procedure for choosing the penalty parameter in linear regression and generalized linear model settings. This permutation procedure was introduced previously in a more limited context in Valdar et al. (2012), and was motivated by work of Ayers and Cordell (2010). The procedure considers the minimal level of penalization required to remove all predictors from the LASSO model under multiple random permutations of the response. Permutation of the response provides a baseline under which there are no true associations with the predictors. Applying a comparable level of penalization to the true (unpermuted) response ensures that the variables included in the model have a joint relationship with the response that is stronger than joint relationships arising by chance.

In recent work, Lockhart et al. (2013) proposed a covariance test for the LASSO that can be applied to parameter selection when variable selection is of primary interest. The test yields a p-value for each model change (inclusion or exclusion of a variable) in the LASSO path. For simple linear models, the test requires that the error variance be known or estimated from the data, limiting general use of the test in this case to data for which the number of samples is larger than the number of predictors. Estimation of the error variance in high dimensional settings, and application of the test to selection of the LASSO penalty parameter are currently under investigation by Lockhart et al..

The remainder of the paper is organized as follows. In the next section, we outline the general data setting and describe the permutation-based selection procedure. In Section 3 we draw some connections between the proposed permutation selection method and the penalty parameters assumed by existing asymptotic theory for the LASSO using recent work on maximal correlations of random vectors on the unit sphere. Section 4 is devoted to a simulation study in which we compare the proposed permutation method with CV, BIC, and a selection procedure derived from the test of Lockhart et al. (2013). In Section 5, we examine the application of different penalty parameter selection methods to multiple real data sets. We conclude with a

brief discussion in Section 7.

2 Permutation Selection of the Penalty Parameter

2.1 Data Setting and Model

We assume that the available data is in the form of an $n \times p$ data matrix \mathbf{X} and an $n \times 1$ response vector \mathbf{y} . The rows of \mathbf{X} correspond to samples; the columns $\mathbf{x}_1, \dots, \mathbf{x}_p$ of \mathbf{X} correspond to measured variables of interest, and are assumed to be standardized. For convenience we describe the proposed permutation scheme in settings where the application of a generalized linear model (GLM) is augmented by use of a 1-norm penalty to derive a sparse model from the available data.

A GLM consists of three components: a random component associated with the response; a systematic component equal to a linear function of the data, and a link function connecting the random and systematic components. Let $\mathbf{y} = y_1, \dots, y_n$ be independent observations of a response variable \mathbf{Y} whose distribution belongs to an exponential family. Here we focus on the Gaussian for continuous responses and on the Binomial for binary, case/control type responses. Let $\boldsymbol{\mu} = E(\mathbf{Y})$ and let $g(\cdot)$ be a known link function. The response and data matrix are linked by the GLM

$$g(\boldsymbol{\mu}) = \mathbf{X}^T \boldsymbol{\beta}. \quad (1)$$

The LASSO procedure (Tibshirani, 1996) fits a sequence, or path, of models characterized by a positive penalty parameter λ that trades off between the overall fit of the model and its complexity. Let $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$ be the 1-norm of $\boldsymbol{\beta}$, and let

$$\ell(\boldsymbol{\beta} : \mathbf{y}, \mathbf{X}) = \sum_{i=1}^n \log L(\boldsymbol{\beta} : y_i, \mathbf{x}_i), \quad (2)$$

be the log-likelihood of $\boldsymbol{\beta}$ given \mathbf{X} and \mathbf{y} , where $L(\cdot)$ is the likelihood function of a single observation under the GLM (1). For a specified value of $\lambda \geq 0$ the LASSO procedure identifies the (unique) model defined by the coefficient vector

$$\hat{\boldsymbol{\beta}}(\lambda : \mathbf{y}, \mathbf{X}) = \arg \max_{\boldsymbol{\beta}} \{ \ell(\boldsymbol{\beta} : \mathbf{y}, \mathbf{X}) - \lambda \|\boldsymbol{\beta}\|_1 \}. \quad (3)$$

For each $\lambda \geq 0$ let

$$S_0(\lambda) = \{j : \hat{\beta}_j(\lambda : \mathbf{y}, \mathbf{X}) \neq 0\} \quad (4)$$

be the set of variables included in the model associated with the coefficient vector $\hat{\boldsymbol{\beta}}(\lambda : \mathbf{y}, \mathbf{X})$. As noted in the introduction, use of the 1-norm penalty ensures that when λ is large the coefficient vector $\hat{\boldsymbol{\beta}}(\lambda : \mathbf{y}, \mathbf{X})$ will

be sparse, or equivalently, the set $S_0(\lambda)$ of selected variables will be small. Use of the LASSO and related procedures arises from practical interest in deriving models of the data that include a relatively small number of selected variables.

2.2 Permutation Selection Procedure

For any permutation π of $[n] := \{1, \dots, n\}$ let $\mathbf{y}_\pi = (y_{\pi(1)}, \dots, y_{\pi(n)})^T$ be a re-ordered version of the response \mathbf{y} . Suppose that permutations π_1, \dots, π_N are obtained by sampling uniformly at random from the set of permutations on $[n]$. For each $1 \leq l \leq N$ and each $\lambda \geq 0$ let

$$S_l(\lambda) = \{j : \hat{\beta}_j(\lambda : \mathbf{y}_{\pi_l}, \mathbf{X}) \neq 0\}$$

be the set of variables selected by the LASSO procedure (3) with response \mathbf{y}_{π_l} and penalty parameter λ . For each random permutation π_l let

$$\lambda_0(\mathbf{y}_{\pi_l}) = \min \{\lambda : |S_l(\lambda)| = 0\} \quad (5)$$

be the smallest value of the penalty parameter for which no variables are selected for the fitted model. Permutation of the response ensures that there is, on average, no systematic relation between \mathbf{y}_{π_l} and the measured predictors in \mathbf{X} . In this case, exclusion of variables from the fitted model correctly reflects the absence of a relationship between \mathbf{y}_{π_l} and \mathbf{X} . The quantity $\lambda_0(\mathbf{y}_{\pi_l})$ is the smallest amount of penalization for which this null relationship is maintained. Our proposed choice of penalty parameter is the median of the observed minimum penalties, namely,

$$\hat{\lambda}_{\text{perm}} = \text{median}(\lambda_0(\mathbf{y}_{\pi_1}), \dots, \lambda_0(\mathbf{y}_{\pi_N})). \quad (6)$$

As can be seen from its definition, the choice of $\hat{\lambda}_{\text{perm}}$ is targeted towards the goal of *variable selection*; the predictive performance of $\hat{\lambda}_{\text{perm}}$ is considered only briefly in the real data analyses below.

Variability of $\hat{\lambda}_{\text{perm}}$, arising from the use of random permutations, can potentially lead to variability in the selected model $S_0(\lambda_{\text{perm}})$. Variability of $\hat{\lambda}_{\text{perm}}$ depends on the number of permutations N , as well as the means by which the values $\lambda_0(\mathbf{y}_{\pi_l})$ are aggregated. Aggregation using the median provides a relatively stable estimate for low to moderate values of N . Ayers and Cordell (2010) proposed a similar penalty selection procedure that uses the maximum of $\lambda_0(\mathbf{y}_{\pi_l})$ over 25 permutations of the response. They showed empirically that the procedure controls the family-wise error rate of falsely including a variable unrelated to the response. However, use of the maximum increases variability of the selected model. In particular, when evaluating λ multiple times on the same data, multiple models can be selected. To reduce this variability, the number of

permutations can be increased, but this has the effect of changing the family wise error rate. In the selection procedure proposed here, increasing the number of permutations yields a better estimate of the true median $\lambda_0(\mathbf{y}_\pi)$, and reduces permutation variability.

In our experience with simulated and real data, $N = 100$ permutations is sufficient to control permutation variability and provide reasonable values of $\hat{\lambda}_{\text{perm}}$. In some cases, a lower N can be sufficient: in Valdar et al. (2012) the median of $N = 20$ permutations was enough to select a stable model $S_0(\hat{\lambda}_{\text{perm}})$ in a genetics setting.

3 Some Connections with LASSO Theory

In the case of linear regression, recent results on maximal correlations provide an asymptotic connection between the permutation based selection procedure described here and existing theoretical work on the consistency of the LASSO. Assume for the moment that the data matrix \mathbf{X} and response \mathbf{y} are related by the standard linear regression model $\mathbf{y} = \mathbf{X}^T \boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a vector of independent mean zero Gaussian errors with common variance. Suppose that \mathbf{y} and the columns $\mathbf{x}_1, \dots, \mathbf{x}_p$ of \mathbf{X} have been centered and scaled to have mean zero and total sum of squares one; then \mathbf{y} and each \mathbf{x}_j lie on the unit sphere S^{n-1} in \mathbb{R}^n . Recall that the standard LASSO coefficient vector is given by

$$\hat{\boldsymbol{\beta}}(\lambda : \mathbf{y}, \mathbf{X}) = \arg \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i^T \boldsymbol{\beta} - y_i)^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}. \quad (7)$$

It is known (cf. Friedman et al. (2010)) that the minimal penalty under which all coefficients of $\boldsymbol{\beta}$ are zero is given by

$$\lambda_0(\mathbf{y}) = \max_{1 \leq j \leq p} |\mathbf{x}_j^T \mathbf{y}|, \quad (8)$$

Thus for any permutation π_l of the response \mathbf{y} ,

$$\lambda_0(\mathbf{y}_{\pi_l}) = \max_{1 \leq j \leq p} |\mathbf{x}_j^T \mathbf{y}_{\pi_l}| \quad (9)$$

is simply the maximum (unsigned) inner product between \mathbf{y}_{π_l} and the columns of \mathbf{X} .

In recent work, Zhang (2013) shows that if both n and p tend to infinity then, letting \mathbf{U} be a random vector uniformly distributed on the unit sphere S^{n-1} ,

$$\inf_{\mathbf{v}_1, \dots, \mathbf{v}_p \in \mathbb{R}^n} \mathbb{P} \left(\max_{1 \leq j \leq p} |\mathbf{v}_j^T \mathbf{U}| \leq \sqrt{1 - p^{-2/(n-1)}} \right) \rightarrow 1. \quad (10)$$

A corresponding lower bound holds if the vectors \mathbf{v}_i are independent and uniformly distributed on S^{n-1} . Under the standard assumption that n and p grow in such a way that $\log p/n \rightarrow 0$, it is easy to see that $\sqrt{1 - p^{-2/(n-1)}}$ is approximately $\sqrt{2 \log p/n}$, in the sense that the ratio between these quantities tends to one as n, p tend to infinity. Thus for each $\delta > 0$, if n and p are sufficiently large and $\log p/n$ is close to zero, then with high probability

$$\max_{1 \leq j \leq p} |\mathbf{x}_j^T \mathbf{U}| \leq (1 + \delta) \sqrt{\frac{2 \log p}{n}}. \quad (11)$$

We expect that a corresponding lower bound, with a different leading constant, will hold for data matrices whose columns \mathbf{x}_j are weakly dependent. Comparing the last display with equation (9), we see that in permutation selection the permuted responses \mathbf{y}_{π_l} act as a proxy for a uniform sample from the unit sphere S^{n-1} , and therefore $\hat{\lambda}_{\text{perm}}$ acts as an estimate of the population quantity

$$\text{median} \left(\max_{1 \leq j \leq p} |\mathbf{x}_j^T \mathbf{U}| \right). \quad (12)$$

The quantity $\sqrt{2 \log p/n}$ appearing in (11) is, up to constants, the value of the penalty parameter assumed in standard results (cf. Bühlmann and van de Geer (2011)) concerning the asymptotic properties of the LASSO for prediction and variable selection. Recent theoretical work (c.f. Hebiri and Lederer (2013), Dalalyan et al. (2014), and the references therein) has examined in some detail how correlations among the columns \mathbf{x}_j of the design matrix \mathbf{X} affect the performance of the LASSO and the selection of the penalty parameter λ . The broad conclusion of this work is that smaller values of λ are appropriate when the columns of \mathbf{X} are more correlated, a relationship that holds for the permutation selection procedure. Indeed, the permutation selection procedure adapts to linear dependence among the columns of the design matrix in a direct way, by assessing their maximum correlation with a permuted response.

To examine the extent to which permutations of the response behave like a uniformly distributed vector on the sphere, we generated a $100 \times 10,000$ data matrix \mathbf{X} with independent standard Gaussian entries. We then computed $\lambda_0(\mathbf{z}_i)$ for each of 10,000 uniformly distributed vectors on S^{n-1} , and compared the density of the resulting values to the values $\lambda_0(\mathbf{y}_{\pi_l})$ from 10,000 permutations of a normalized response \mathbf{y} in two settings: a sparse setting with $s = 10$ true variables and a signal to noise ratio (SNR, defined in (14)) equal to 2; and a less sparse setting with $s = 1000$ true variables and SNR equal to 2. The results, shown in Figure 1, demonstrate good agreement in both cases with the penalties derived from independent uniform responses. Simulation results for data matrices \mathbf{X} with different correlation structures (not shown) were similar.

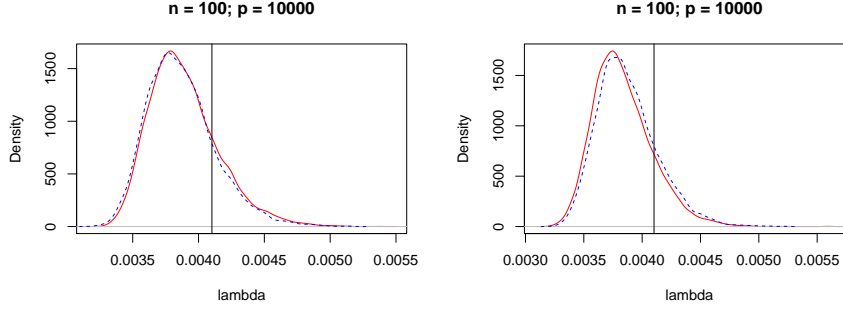


Figure 1: Distributions of $\lambda_0(\tilde{\mathbf{y}}_\pi)$ (red) and $\lambda_0(\tilde{\mathbf{z}})$ (blue) for $n = 100$ and $p = 10000$. (Left) sparse model with $s = 10$. (Right) non sparse model with $s = 1000$. The black vertical line indicates the value of $\sqrt{1 - p^{-2/(n-1)}}$.

4 Simulations

We performed a simulation study in which the proposed permutation selection procedure and competing methods were used to select the penalty parameter of the LASSO. Competing methods were BIC, 10-fold CV, and a procedure based on the covariance test of Lockhart et al. (2013). Simulated data sets differed in their dependence and level of sparsity (number of truly active predictors). Both linear (Gaussian) and logistic (binomial) settings were considered.

4.1 Predictor Matrix Generation

We simulated predictor matrices with independent, multivariate Gaussian rows (samples). In particular, the n rows of \mathbf{X} were generated as independent samples from a $\mathcal{N}_p(\mathbf{0}, \Sigma)$ distribution. Several structures for the covariance matrix Σ were considered.

- (A) Independent: $\Sigma = \mathbf{I}_p$;
- (B) Block correlation structure: $\sigma_{ij} = 0.5$ if $i \bmod 10 = j \bmod 10$ and $\sigma_{ij} = 0$ otherwise;
- (C) Fast decaying AR(1) correlations: $\sigma_{ij} = 0.9^{|i-j|}$;
- (D) Slowly decaying AR(1) correlations: $\sigma_{ij} = 0.99^{|i-j|}$.

4.2 Response Generation

Given the predictor matrix \mathbf{X} and a simulated effects vector β (for more details, see below), we modeled the response using the generalized linear model (GLM) in equation 1. Two GLM settings were considered: Gaussian (standard linear model) and Bernoulli (logistic regression).

4.2.1 Linear Regression Model

Gaussian regression responses were simulated based on the standard linear regression version of equation (1), in which the link function $g(\cdot)$ is the identity. Given a set of s true variables selected uniformly at random, we generated the components of the associated effect vector β_s independently from a $\beta_i \sim U(0.25, 1)$ distribution. Letting \mathbf{X}_s be the restriction of the predictor matrix to the columns of the s selected variables, the response vector \mathbf{y} was generated as

$$\mathbf{y} = \mathbf{X}_s^T \beta_s + \epsilon \quad (13)$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ is a vector of independent Gaussian errors. The variance σ^2 was selected to achieve a desired signal to noise ratio (SNR) given by

$$\text{SNR} = \sigma^{-2} \beta_s^T \text{Var}(\mathbf{X}_s) \beta_s. \quad (14)$$

One hundred simulations were performed for each choice of the following simulation parameters (64 in total):

- sample correlation structures (A)-(D);
- $\text{SNR} = 0.5, 2$;
- number of subjects $n = 200, 1000$;
- number of true variables $s = 1, 5, 10, 20$.

In each case, the number of variables p was equal to 500.

4.2.2 Logistic Regression Model

Logistic regression responses were simulated based on Bernoulli draws from the logistic version of the generalized linear model (1) with link function

$$\text{logit}(q) = \log\left(\frac{q}{1-q}\right), \quad (15)$$

where q is the probability that a given subject is a case. In particular, for each sample i we generated a value

$$q_i = \mu + \mathbf{x}_{i,s}^T \beta_s, \quad (16)$$

where $\mu = \mathbf{1}^T (\mathbf{X}_s^T \beta_s)$ is the intercept needed to obtain expected balance in the number of cases and controls, $\mathbf{x}_{\cdot,s}$ are the s predictors having true effects, and β_s are the effects, selected independently with $\exp\{\beta_{sj}\} \sim$

$N(\mu_\beta, 0.02^2)$. The phenotype of sample i was drawn from a Bernoulli(q_i) distribution. The parameter μ_β was used to control the signal level in each of the high and low dimensional data settings. For the low dimensional setting, we selected μ_β as $\log(1.15)$ and $\log(1.35)$ for low and high signal levels, respectively. In the high dimensional setting, we selected μ_β as $\log(1.75)$ and $\log(2.5)$ for low and high signal levels, respectively. For each signal level, one hundred simulations were performed for each combination of correlation structure, sample size n , and number of true variables s used in the Gaussian setting.

4.3 Competing λ -selection methods

4.3.1 Bayesian Information Criterion (BIC)

A standard version of BIC was implemented, selecting penalty parameter λ_{BIC} via the relation

$$\lambda_{\text{BIC}} = \underset{\lambda}{\operatorname{argmin}} \left\{ -2\ell(\hat{\beta}(\lambda : \mathbf{y}, \mathbf{X}) : \mathbf{y}, \mathbf{X}) + \text{df}(\lambda) \log(n) \right\}, \quad (17)$$

where $\text{df}(\lambda) = |S_0(\lambda)|$ is the number of non-zero coefficients in the coefficient vector $\hat{\beta}(\lambda : \mathbf{y}, \mathbf{X})$. The resulting variable set is $S_0(\lambda_{\text{BIC}})$

4.3.2 Cross Validation (CV)

Cross-validation was implemented using the `cv.glmnet()` function from the R-package `r/glmnet`, as described in (Friedman et al., 2010). Specifically, we selected the value of λ minimizing the K-fold cross validation error

$$\lambda_{\text{CV}} = \underset{\lambda}{\operatorname{argmin}} \left\{ \sum_{k=1}^K \sum_{i \in V_k} \left[y_i - \mathbf{x}_i^T \hat{\beta}^{(-k)}(\lambda : \mathbf{y}^{(-k)}, \mathbf{X}^{(-k)}) \right]^2 \right\}. \quad (18)$$

Here V_1, \dots, V_K is a randomly chosen partition of $\{1, \dots, n\}$ into groups of size $\lfloor n/K \rfloor$ or $\lceil n/K \rceil$, and the superscript $(-k)$ indicates that the subjects in V_k have been excluded. The resulting variable set is $S_0(\lambda_{\text{CV}})$. We report results based on the value $K = 10$. The results for $K = 3$ and $K = n$ (also referred to as the jackknife) were similar.

4.3.3 Covariance Test

The covariance test of Lockhart et al. (2013) provides a list of p-values corresponding to points in the LASSO path where a variable is added to the model. In principle, these p-values can be used to select a value of the penalty parameter; Lockhart et al. left the specification of such a procedure for future work. Here we propose a simple procedure to select a penalty parameter λ_{CT} using the p-values from the covariance test.

Let p_r be the p-value produced by the covariance test for the r th change in the LASSO path, and let

$$r^* = \max\{r : p_r \leq \alpha\}, \quad (19)$$

where α is a fixed significance level, which we take to be 0.05 in what follows. Define λ_{CT} to be the value of λ at which the r^* th change in the LASSO path takes place; the resulting variable set is then $S_0(\lambda_{CT})$. In the case that no p-values are less than α , we define $S_0(\lambda_{CT}) = \emptyset$.

Use of λ_{CT} is limited to settings in which the covariance test can be applied. At present the covariance test cannot be applied in high dimensional ($p > n$) linear regression settings when the error variance is unknown. Computation of p-values for the covariance test requires knowledge of the LASSO path, in particular, values of λ where model changes take place. Use of the LARS algorithm for this purpose can be problematic: in several of the simulated examples, the LARS algorithm encountered matrix singularities for the active set of variables, and we were unable to exactly identify r^* . In such cases, we identify r^* based on the set p-values available from the portion of the LARS path that was successfully fit. We note that current methods for LASSO fitting such as `r/glmnet` assess the LASSO path at a grid of λ values (e.g., 100 values between λ_0 and $\epsilon\lambda_0$ for epsilon close to 0). This grid evaluation allows all other selection methods considered here to be fully evaluated; but as such grid based procedures are unable to locate the exact change points on the path, they can not be used for the covariance test.

4.4 Simulation Results

Across our simulations, the relationships between penalty selection methods were relatively consistent, with overall performance depending primarily on the complexity of the chosen parameters (stronger dependence, lower SNR, fewer samples, and more active variables). Below we describe the simulation results using the block correlation structure (covariance structure B), which are representative of our overall findings. Figures showing simulation results under other settings can be found in Appendix A.

4.4.1 Gaussian Simulation Results

For the classical ($n > p$) setting we were able to apply all four selection methods. Application of the covariance test to high dimensional ($p > n$) problems when the error variance is unknown is still under investigation. Rather than exclude λ_{CT} from the high dimensional simulations, we evaluated λ_{CT} using the simulated error variances. Thus the covariance test is given a substantial advantage in the high dimensional settings, and the results from these settings should be interpreted accordingly.

Figure 2 shows the average power (number of true discoveries divided by total number of active variables)

and the average false discovery rate (number of false discoveries divided by total number of discoveries) for high and low SNR regimes, and for different numbers of true variables s . Examining the results from the low dimensional data setting (left), we observe a consistent relationship between power and FDR. Specifically, we find that CV has the highest power along with the largest FDR. Permutation selection and BIC perform similarly, with BIC having slightly increased power and FDR. The covariance test is the most conservative of those considered, having both lower power and lower FDR than the other methods. In the high dimensional setting (right), we observe a similar relationship between power and FDR. The principal difference is that BIC and permutation selection have greater differences in power and FDR in the high dimensional setting. Specifically, while BIC's advantage in power over permutation selection is slightly larger in the high dimensional setting, it comes with a greater increase in FDR than that observed in the low dimensional setting.

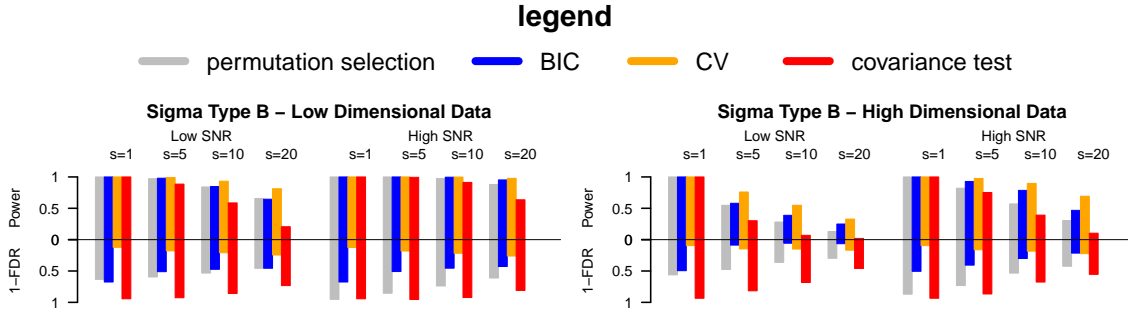


Figure 2: Bar plots of power and false discovery rate (FDR) for selected gaussian regression simulation settings. s indicates the number of true signals.

4.4.2 Logistic Simulation Results

Figure 3 summarizes the results of the logistic simulations. (The covariance test is applicable to logistic regression in both low and high dimensional settings.) Examining the low dimensional data setting (left), we observe that CV has the highest power, but also the highest FDR. Permutation selection is generally comparable, or superior, to BIC. The covariance test performs best for $s = 1$; in the low SNR regime its performance degrades as s increases. Overall, the covariance test is more conservative, with lower power and lower FDR than the other methods. Similar trends are seen with the high dimensional data.

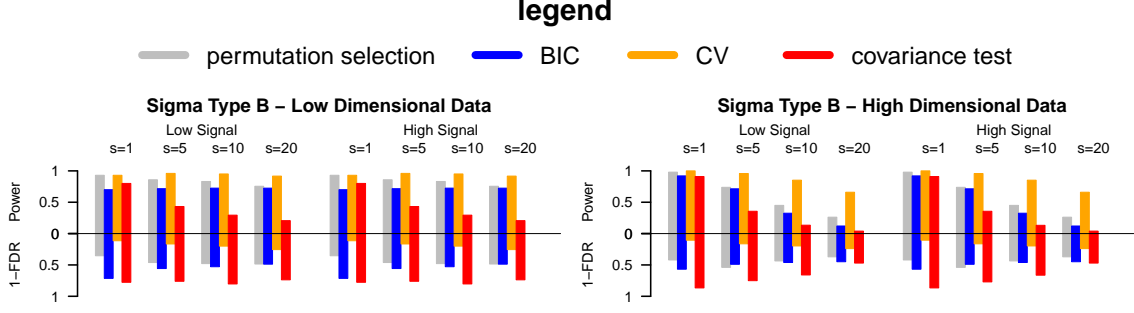


Figure 3: Bar plots of power and false positive rate (FPR) for selected logistic regression simulation setting. s indicates the number of true signals.

5 Real Data Analysis

Here we apply all four considered penalty parameter selection methods on several real data sets. The data sets examined are:

- Internet-Ad data (Kushmerick, 1999): A document classification problem consisting of mostly binary features. The response is binary, and indicates whether the document is an advertisement. Only 1.2% of the values in the predictor matrix are nonzero. The data consist of 2358 observations of 4290 variables. (This data is publicly available at <http://www.stanford.edu/hastie/glmnet/>)
- The Cancer Genome Atlas Network (TCGA) breast cancer data (TCGA, 2012): Tumor and germline DNA samples were obtained from 825 patients. Rsem gene expression values were normalized to the upper quartile of the total counts. The data was then log2 transformed and genes were median centered. Genes were filtered such that at least 70% of samples had a value, and then imputation was performed. (This data is available from the TCGA data portal at <https://tcga-data.nci.nih.gov/tcga/>) We examine the problem of distinguishing 371 Luminal A cancer samples from 170 Luminal B cancer samples based on the expression values of the 17,007 genes that remained after quality control.

We consider the computational time of each fitting procedure and, although our primary focus is variable selection, we also examine the predictive ability of the fitted models. For each analysis, we randomly split the data set into a training and a test set. The training set consists of two thirds of the data, and the remaining third is used for testing. The LASSO was applied to the training data with the penalty parameter selected by each method, and predictive performance was evaluated on the test set. Random splitting of each data set was repeated 10 times.

5.1 Internet Ad Data Analysis

Table 1 summarizes the sizes of the selected models, the percent of the test set misclassified, and computational time for each penalty selection method. When applying the covariance test procedure, we found that collinearity in the data created problems, and we only had access to the first 15 steps in the LASSO path (with some variability for different splits). This has two consequences: (i) the model resulting from the covariance test may be larger than if we had access to the full path; and (ii) the reported computation time for the covariance test is significantly reduced. Examining how often the different procedures misclassified the test set, we found that the prediction based CV method had the best performance, followed by BIC and permutation selection. The more conservative covariance test had the highest misclassification rate. As expected, the simple permutation selection procedure was substantially faster than all other methods.

| Method | Model Size | Percent Misclassified | CPU seconds |
|-----------------------|--------------|-----------------------|--------------|
| BIC | 35.6 (2.01) | 5.07 (0.36) | 12.08 (0.32) |
| covTest* | 6 (0.88) | 9.89 (0.84) | 18.57 (1.91) |
| CV | 123.7 (6.67) | 3.46 (0.23) | 34.35 (0.95) |
| Permutation Selection | 26.5 (2.13) | 6.45 (0.28) | 3.69 (0.04) |

Table 1: Means and standard errors of the model sizes, percent of test set misclassified, and computation times from 10 random splits of the Internet Ad data. * indicates that the exact model may be larger, but LARS was unable to fit the entire path, which also reduced computation time.

5.2 TCGA Data Analysis

Table 2 summarizes the sizes of the selected models, the percent of the test set misclassified, and computational times for each penalty selection method. Collinearity issues arose again in our application of the covariance test, and we therefore limited our analysis to the first 100 steps in the LARS path; as with the Internet Ad data, this significantly lowered the computational time for the covariance test. Examining the percent of the test set misclassified by the models, we find that CV has the best performance, followed by permutation selection and BIC. The covariance test has a significantly higher misclassification rate than the other methods. BIC and permutation selection have similar computational times, closely followed by CV. In this analysis the covariance test was significantly (more than one order of magnitude) slower than the other methods.

| Method | Model Size | Percent Misclassified | CPU seconds |
|-----------------------|-------------|-----------------------|---------------|
| BIC | 18.5 (2.62) | 14.61 (1.28) | 9.43 (0.12) |
| covTest* | 8.4 (5.15) | 26.33 (2.47) | 601.05 (6.72) |
| CV | 85.8 (4.12) | 8.05 (0.68) | 14.15 (0.19) |
| Permutation Selection | 22.9 (0.81) | 13.00 (0.51) | 10.75 (0.26) |

Table 2: Means and standard errors of the model sizes, percent of test set misclassified, and computation times from 10 random splits of the TCGA luminal subsets data. * indicates that the exact model may be larger, but LARS was unable to fit the entire path, which also reduces computation time.

6 Summary

From the simulated and real data analyses, several trends are apparent. As expected, the prediction based cross validation method tends to select larger models from the LASSO sample path. These models have high power, but relatively large false discovery rate; other penalty selection methods are preferred when variable selection is the primary goal of the analysis. In contrast with cross validation, the covariance test tends to select smaller models, with lower power and lower false discover rate. Permutation selection and BIC lie somewhere between cross validation and the covariance test, favoring moderate sized models and a more balanced tradeoff between power and false discovery rate.

The covariance test method performs best, often beating other methods, when the number s of true variables is small, but its performance drops off as the number of true variables increases. This may reflect the difficulty of testing changes in the LASSO path when conditioning on more complex models containing a mix of true and spurious variables. At present, application of the covariance test to high dimensional ($p > n$) linear models is limited to settings in which the error variance is known or can be accurately estimated. In our high dimensional linear model simulations, we provided the true error variance to the covariance test procedure. We note that in cases where the number of true variables is known to be very small, simple predictor-by-predictor selection methods may outperform the LASSO, regardless of how the penalty parameter is selected.

The permutation selection method is straightforward in its implementation and interpretation. In our simulations, permutation selection is comparable or superior to BIC, and is generally competitive with the covariance test, the former having the advantage for large values of s , the latter for small values of s . In the real data analyses, permutation selection and BIC were roughly comparable in terms of model size and predictive performance. The variables selected by the covariance test were also selected by BIC and permutation selection. As it tends to produce smaller models, the covariance test procedure had relatively poor predictive performance.

Cross validation, BIC, and the covariance test require computation of the full LASSO path on the given data. By contrast, permutation selection requires only the initial point of the LASSO sample path, but under

multiple permutations of the response. On balance, the efficiency with which the initial point of the LASSO path can be identified makes permutation selection faster, sometimes by an order of magnitude, than the other methods considered.

7 Discussion

When selecting models using penalized regression methods such as the LASSO, selection of the penalty parameter is an important part of the analysis process. Selection is often done with the goal of prediction or variable selection. The most common methods for selecting the penalty parameter are cross validation (CV) and the Bayesian information criterion (BIC). In this paper we introduced a simple permutation based selection procedure that is intended for situations where variable selection is a primary goal of model fitting. Permutation selection chooses a model whose variables have a joint relationship with the response that is stronger than joint relationships occurring at random, for permuted versions of the response.

Permutation selection is fast and yields interpretable models. In our simulations and real data analyses, permutation selection was comparable to BIC and competitive with the covariance test procedure. At present, application of the covariance test to high dimensional ($p > n$) linear models is limited to settings in which the error variance is known or can be accurately estimated. Although we have focused here on the LASSO, permutation selection is applicable in principle to other penalized regression methods such as SCAD.

We have considered settings in which all variables are subject to penalization. In some cases it is natural to consider models in which select variables (for example, predictors known to affect the response) are unpenalized. Although CV and BIC can be extended to this situation, permutation selection does not immediately apply, as permutation of the response will nullify its relationship between penalized and unpenalized variables alike. We leave to future work a detailed investigation of how permutation selection can be applied to applications with penalized and unpenalized variables.

Both permutation selection and cross validation make use of randomization, which leads to variability in the selection of the penalty parameter. For permutation selection this variability is controlled by considering the median of the null penalties across a moderate number (100) of permutations. In most applications with moderate or small sample sizes, the randomness of the sample itself is a more significant source of variability, one that is often not accounted for in applications of the LASSO, regardless of how the penalty parameter is selected. Recently, a number of resampling based methods have been proposed to address the stability of models with respect to variability of the sample. We refer the interested reader to Valdar et al. (2009), Meinshausen and Bühlmann (2010), and Valdar et al. (2012) for more details.

A Full Results Plots

A.1 Gaussian Model Results

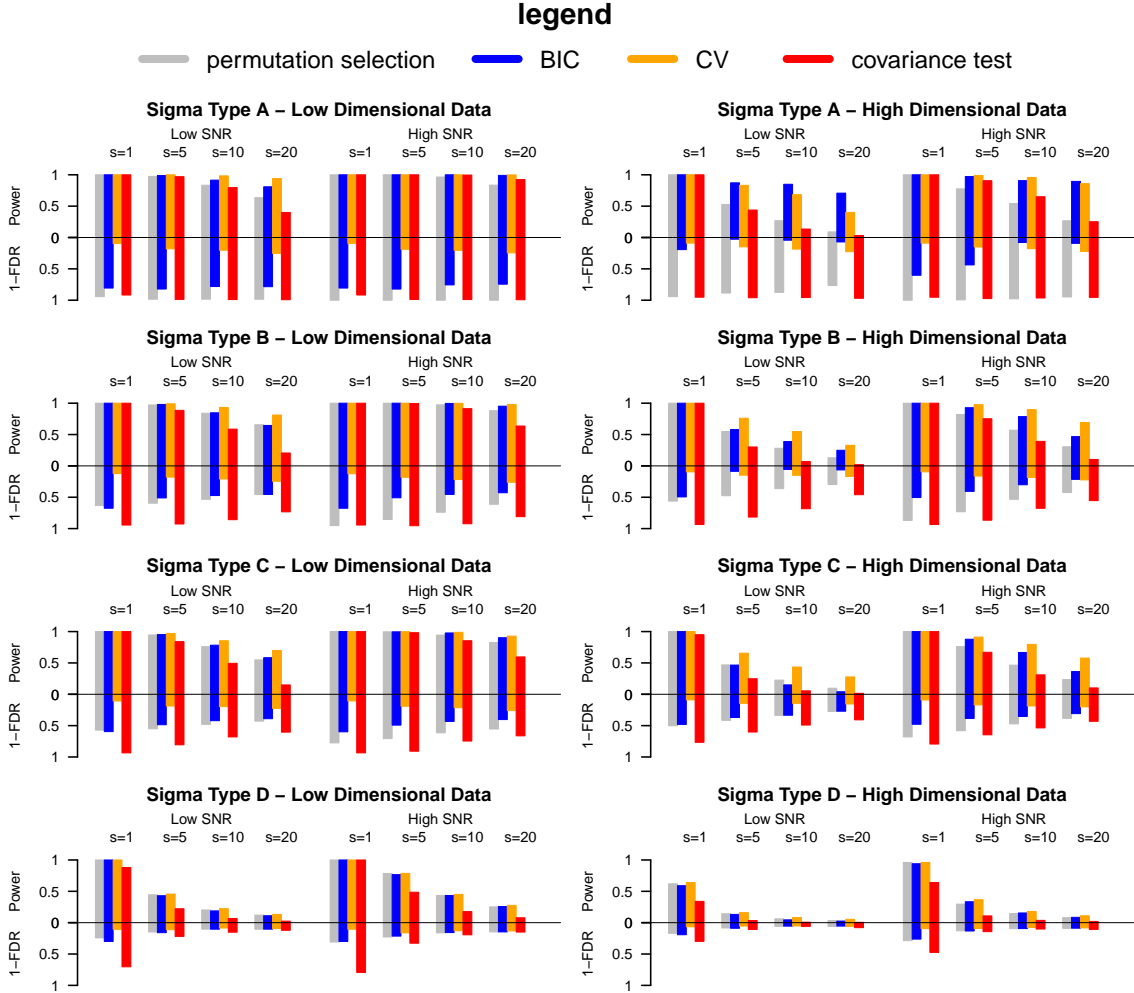


Figure 4: Bar plots of power and 1– false discovery rate (FDR) for each gaussian regression simulation setting.

A.2 Logistic Model Results

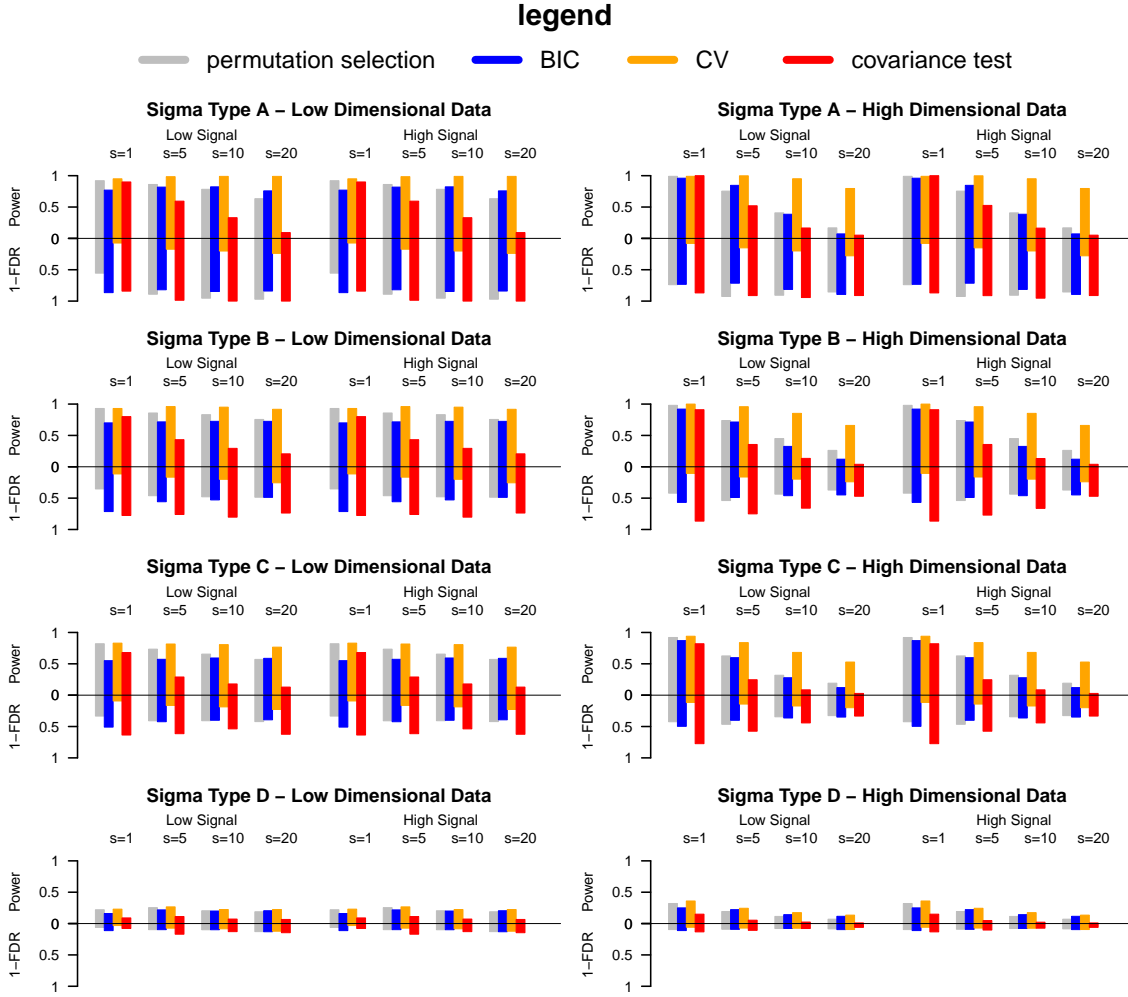


Figure 5: Bar plots of power and 1– false discovery rate (FDR) for each logistic regression simulation setting.

References

- Ayers KL, Cordell HJ. 2010. SNP Selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic epidemiology* 34:879–91.
- Bühlmann P, van de Geer S. 2011. *Statistics for High-Dimensional Data - Methods, Theory and Applications*. Springer Series in Statistics 0172-7397. Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-20192-9.
- Dalalyan AS, Hebiri M, Lederer J. 2014. On the Prediction Performance of the Lasso. *ArXiv e-prints* .

- Fan J, Li R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96:1348–1360.
- Feng Y, Yu Y. 2013. Consistent Cross-Validation for Tuning Parameter Selection in High-Dimensional Variable Selection. *ArXiv e-prints* .
- Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33:1.
- Hebiri M, Lederer J. 2013. How correlations influence lasso prediction. *IEEE Transactions on Information Theory* 59:1846–1854.
- Kushmerick N. 1999. Learning to remove Internet advertisements. *ACM Press*. p 175–181.
- Lockhart R, Taylor J, Tibshirani RRJ. 2013. A significance test for the lasso. *arXiv preprint arXiv: 1308.4004*. :1–44.
- Meinshausen N, Bühlmann P. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72:417–473.
- TCGA. 2012. Comprehensive molecular portraits of human breast tumours. *Nature* 490:61–70. doi:10.1038/nature11412.
- Tibshirani R. 1996. Regression shrinkage and selection via lasso. *Journal of the Royal Statistical Society Series B (Methodological)* 58:267–288.
- Tibshirani R. 2011. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73:273–282. doi:10.1111/j.1467-9868.2011.00771.x.
- Valdar W, Holmes CC, Mott R, Flint J. 2009. Mapping in structured populations by resample model averaging. *Genetics* 182:1263–77.
- Valdar W, Sabourin J, Nobel A, Holmes CC. 2012. Reprioritizing genetic associations in hit regions using lasso-based resample model averaging. *Genetic epidemiology* 36:451–62.
- Zhang K. 2013. Rank-Extreme Association of Gaussian Vectors and Low-Rank Detection. *ArXiv e-prints* .